Joint USDA-DOE Plant Genomics Knowledgebase Workshop Report

Setting the Stage for the Plant Knowledgebase Workshop: Bioinformatics Use in Advancing Plant Genomics, Genetics, and Breeding

Friday, January 8, 2010, 10:30 a.m. – 4:00 p.m.
Plant and Animal Genome XVIII
Town & Country Hotel
San Diego, California

Convened by the

U.S. Department of Energy (DOE)
Office of Science
Office of Biological and Environmental Research

U.S. Department of Agriculture (USDA)
National Institute of Food and Agriculture

Workshop Organizers: Catherine Ronning (DOE), Susan Gregurick (DOE), Ed Kaleikau (USDA), Gera Jochum (USDA), and Bob Cottingham (Oak Ridge National Laboratory)

Audience: 100 plant scientists, geneticists, breeders, and bioinformatics specialists

Speakers: Catherine Ronning (DOE Office of Biological and Environmental Research), Bob Cottingham (Oak Ridge National Laboratory), David Francis (Ohio State University), Steve Rounsley (University of Arizona), Eva Huala (The Arabidopsis Information Resource), Doreen Ware (Cold Spring Harbor Laboratory), and Dan Rokhsar (DOE Joint Genome Institute)

Introduction

The Department of Energy (DOE) Genomic Science program supports systems biology research to ultimately achieve a predictive understanding of microbial and plant systems for advancing DOE missions such as sustainably producing biofuels, investigating biological controls on carbon cycling, and cleaning up contaminated environments. To manage and effectively use the exponentially increasing volume and diversity of data resulting from its projects, the Genomic Science program is developing the DOE Systems Biology Knowledgebase (genomicscience.energy.gov/compbio/).

A DOE workshop held in May 2008 defined the vision for the Knowledgebase—an open cyberinfrastructure to integrate systems biology data, analytical software, and computational modeling tools that will drive two classes of work: (1) experimental design and (2) modeling and simulation. This community-driven Knowledgebase will need to be understandable and accessible to the entire research community and must have an intuitive design that facilitates sharing and contribution among all users. To provide computational capabilities that support DOE systems biology research and other application areas, the Knowledgebase will need to serve multiple roles, including a flexible, adaptable repository of data and results from high-throughput experiments; a collection of tools to derive new insights through data synthesis, analysis, and comparison; a framework to test scientific understanding; a heuristic capability to improve the value and sophistication of further inquiry; and a foundation for prediction, design, manipulation, and, ultimately, engineering of biological systems.

For the USDA National Institute of Food and Agriculture, a grand challenge in plant genomics, genetics, and breeding is to identify gene combinations that lead to significant innovation in agriculture and production of raw materials for food, feed, fiber, and fuel. An interdisciplinary approach such as molecular plant breeding may be able to meet this challenge and revolutionize 21st century plant improvement. Molecular plant breeding is founded on the integration of advances in biotechnology, genomic research, and molecular marker applications with conventional plant breeding practices. This integration would require a combination of molecular markers and high-throughput genome sequencing efforts, new knowledge of genome structure and function, statistical approaches to estimate genetic effects, experience in both laboratory molecular methods and field-based breeding practices, and the ability to manage large datasets with diverse data types. This workshop also was intended to assist in developing strategies to expand bioinformatic tools to enable the breeder-centric, high-throughput data management and visualization tools and platforms necessary for integrating genome sequence information with other data types and to provide the breeder-centric views of map and trait data that best serve plant breeders' needs. Implementing such strategies will require (1) broadly training a new generation of plant researchers to fully master key areas such as bioinformatics and quantitative genetics and breeding; (2) establishing partnerships with universities, federal laboratories, industry, and international centers to take advantage of the best training opportunities; and (3) developing a new cohort of researchers able to translate and integrate basic research endeavors with applied plant improvement and value added outcomes for sustainable bioenergy production systems.

Workshop Description

The Plant Genomics Knowledgebase Workshop—held in conjunction with the Plant and Animal Genome XVIII conference in San Diego, California—brought together 100 plant scientists, geneticists, breeders, and bioinformatic specialists to discuss current issues facing plant breeders in light of ever-increasing amounts of genomic data. The workshop featured lectures by leaders in the plant breeding, genomics, and bioinformatics communities. These presentations set the stage for afternoon breakout discussions by addressing the data needs of more-applied breeding programs and describing resources emanating from more-fundamental plant genomics and bioinformatics research. This event is part of a series of DOE-supported workshops to engage the scientific community in discussing scientific objectives the Knowledgebase could serve and determining which endpoints could be achieved in the near, mid-, and long term.

The overarching goal of the workshop was to address the following question:

> How can we best design the Knowledgebase to have the flexibility to grow with and adapt to new data and information challenges in the future?

A key objective was to specifically identify the requirements for effectively developing data capabilities for systems biology as applied to plants, particularly the research and development of plant feedstocks for biofuels. The current state of plant informatics is represented by many disparate databases primarily focusing on specific taxonomic groups or processes. To enable a systems biology approach to plant research, integrating all types of data (including molecular, morphological, and "-omics") for bioenergy-relevant plant species is important. Thus, the challenge will be to develop uniformity of data format and database architectures to effectively integrate diverse data types and enable user-friendly acquisition and analysis.

Charge Questions

All participants were asked to address two charge questions:

1. What types of experimental data are currently available, and of these, which format(s) are most useful and valuable? Can data from various sources and of various types be standardized into this "ideal" format and then be organized and integrated into one common, searchable application?

   For example, a researcher studying cell wall biosynthesis in grasses may benefit from work being performed in poplar. How can we best facilitate cross-species comparisons? How can we use these tools to leverage and apply knowledge gained from model species (e.g., *Arabidopsis* and rice) to crop plants?

2. What are the challenges for plant bioinformatics in a 2- to 3-year time frame? Given the development of an integrated, uniform system (Question 1), what types of analyses do you foresee, and what types of analysis tools will maximize the system's utility?

   How do we best organize, for example, pathways and processes, and how can we organize and distinguish common processes from those that are taxon-specific? How can these informatics resources best be used to enhance plant breeding (i.e., "genotype to phenotype")? Will these resources be effective in designing decision support tools for plant breeders in the field?

Summary of Workshop Recommendations

Three recommendations from the workshop are:

1. **Establish community–agreed upon data formats and storage protocols for environmental and experimental metadata and workflows.**

   This includes gene annotations; gene product functions; protein interactions; expression and methylation data; natural variation data; and phenotypic data such as geographical coordinates of a field, sampling dates, weather conditions, experimental designs, scoring methods, and images. Although some of these data types and metadata informatics have well-established formats and protocols, others do not and are not well linked to upstream genomic data. Standards development endorsed by the research community needs to be a collaborative and iterative effort between data generators and developers of cyberinfrastructure such as the Knowledgebase. Active, community-driven development of standards will require resource commitments in the form of coordination workshops; new tools to facilitate annotation and data deposition; curation; and compliance through journals, agencies, and peers.

2. **Develop the ability for comparative analyses of gene sequences, transcript and protein abundance, phenotypes, and the relationships among these components across multiple species.**

   Because the plant community comprises both systems biologists and plant breeders, the Knowledgebase must be adaptive to different user needs. Developing comparative analyses across species will require different levels of community support for different research needs. Moreover, coordination is needed among the various plant bioinformatic efforts sponsored by different agencies to avoid duplication of effort and to identify opportunities for collaboration.

3. **Establish long-term support for maintaining repositories of a variety of genomic and phenotypic data types.**

   This will be key to success of a knowledgebase that tries to integrate information from these resources.

Data and Analytical Challenges for Bioenergy Feedstocks

Although this workshop focused on data capabilities relevant to developing plant feedstocks for biofuels, many of the tools, approaches, and issues discussed are applicable to non-biofuel plant species, including well-studied model organisms such as *Arabidopsis*. Thus, Knowledgebase efforts can be leveraged to other plant bioinformatics systems and biological research areas and vice versa. Workshop participants identified several data and analytical issues for plant genomics, including the diversity of data types available, the challenges of dealing with phenotypic data, cross-species analyses, data integration, and standards for interoperability among data and information resources.

Available Data Types for Plant Genomics

The range and quality of available data depend on the extent to which a particular genome has been studied. For a well-studied model organism like *Arabidopsis*, a broad range of data types supported by a rich history of published research helps researchers move from gene sequence to molecular function, associated phenotype, relevant metabolic or regulatory pathways, and interaction partners. As the types of data being generated for different species of bioenergy feedstocks continue to grow, a top priority will be developing appropriate repositories for handling each data type.

What Kinds of Data are Available from Arabidopsis Research?

- **High-quality genome annotation.** The annotated genome forms the basis of all other "-omics" data. The *Arabidopsis* genome has been revised nine times since the initial sequence was completed in 2000, and its annotation continues to evolve. Current revisions to the annotation include adding splice variants and untranslated regions (i.e., 5' and 3' UTRs) as transcript data improves, correcting sequencing errors, and adding features that are more difficult to annotate such as noncoding RNAs and genes that encode small proteins. In the last 5 years, The Arabidopsis Information Resource (TAIR) has added or updated about half of the genes in the current release, and large, new datasets continue to be generated. Revising genome annotation is a continuous process.

- **Experimental gene function data.** In addition to refining gene structures, TAIR curators have been adding gene function annotations based on experimental data from research articles. To date, 8,622 genes have been annotated with results from published experiments—a total that continues to increase rapidly. Most data used in this manual annotation process were not from high-throughput experiments but from those focusing on a single gene. Gene Ontology annotations describe biological process, molecular function, and cellular compartment. Plant ontology annotations describe the anatomical part and developmental stage associated with expression patterns. This is a rich dataset to consider transferring to other plant species.
- **Phenotypic data** for *Arabidopsis* have largely been qualitative. Currently, these data are in a free-text form, and efforts are needed to use a plant ontology to describe these phenotypes. *Arabidopsis* phenotypic data also include about 5,000 images in a form that is not yet readily transferable to other plants.
- **Protein interaction data** build on existing foundational datasets to generate networks of interactions.
- **Natural variation data** include more quantitative data than some other kinds of data.
- **Expression, methylation, pathways, and networks data** provide more of a genome-wide view of how this plant functions.

Next-Generation Sequencing Data. With the expanding use of next-generation sequencing technologies such as Illumina and 454, an important challenge will be dealing with the vast volume and variable quality of short read sequences generated by many different sources. Needed are resources for assembling and curating these massive amounts of data and tools for using the data to identify and develop single nucleotide polymorphism (SNP) markers, such as the current iPlant effort.

Environmental Metadata. One of the more difficult data challenges identified by workshop participants will be defining appropriate data formats and storage protocols for environmental and experimental metadata. Such data include geographical field coordinates, sampling dates, weather conditions, experimental design, scoring methods, and images. Metadata collection systems will need to be standardized and automated (e.g., using bar codes).

Phenotypic Data Challenges

Enabling large-scale generation of useful phenotypic data and ensuring easy access to it are some of the most important challenges for the bioenergy feedstock research community. Many bioinformatic efforts for plant biology have emphasized non-phenotypic data (e.g., DNA sequence, SNP markers, gene expression, and epigenetics). Phenotypic data—an extremely broad category of data—are subject to considerable noise, have few or no uniform standards, and are highly dependent on genetic context (e.g., particular individuals that have a specific genotype) and environmental context (e.g., timescales, locations, and precipitation). Some critical challenges identified by workshop participants include developing standards and more efficient methods for generating and managing phenotypic data, improving the ability to link specific genes to phenotypes on a quantitative scale, and establishing central repositories for storing phenotypic data and genetic material. One organization that will address these issues is the International Plant Phenomics Initiative (www.plantphenomics.com), which is being organized by European, Canadian, and Australian researchers to promote international collaboration for plant phenomics.

Limited Availability of Phenotypic Data. The availability of phenotypic data is key to identifying quantitative trait loci and genes associated with important bioenergy-related traits. However, phenotypic data is currently limited. At present, for example, 33% of the protein-coding genes in the *Arabidopsis* genome have experimental annotations, and only 9% have phenotypic descriptions (including “no visible phenotype”). Understanding of genes in biofuel species is even less developed. The lack of robust phenotypic data and functional annotation will result in the continued extensive use of transitive annotation based on sequence similarity from generic databases such as Pfam and UniProt. This is a primitive approach to improving our understanding of plant biology with respect to biofuels.

The amount of meaningful phenotypic data available in public databases is very small compared to the amount of genomic data available. Moreover, the limited resources handling phenotypic data do not address all phenotypes and often do not include lines used for breeding. They thus are not providing breeders with needed information. Participants suggested the need to develop a system of quality scores that could provide a measure of confidence for the heritability and/or measurement of a particular phenotype.

For many applied breeding objectives, a greater focus is needed on generating more phenotypic information in more populations of a given species and, importantly, generating data in actual elite breeding populations. Collecting phenotypic data for complex traits in plants is time consuming. Many potential bioenergy crops are perennial, so successive-year data are needed for individual plants or accessions—information difficult and expensive to obtain. In addition, measuring environmental effects on phenotypes, which requires quantitative data, is as important as defining genotype.

More Objective and Quantitative Phenotypic Data. Descriptors supported by the Union for the Protection of New Varieties of Plants (UPOV) or USDA’s National Plant Germplasm System (NPGS) Germplasm Resources Information Network (GRIN) are used by breeders to classify traits into defined categories. For example, in GRIN, a trait such as color is assigned a numerical color category like “1” for green or “2” for yellow. Although these descriptor systems attempt to make all trait data more uniform, they fail to account for inherent variation within an accession (e.g., how “green” is it?). They also are disassociated from contemporary systems of measurement and disconnected from data scales used by expert practitioners.

Trait data should be quantitative and objective whenever possible. For example, there are very objective systems for measuring color, such as the RGB system for computers. High-throughput systems are needed that can extract quantitative phenotypic data from images. The advantages of such objective measures for phenotype are clear: the ability to interconvert systems of measurement and the ability to easily obtain estimates of variation (and therefore estimates of heritability). Tools that enable the mapping of one system to another (e.g., a descriptor to an ontology and a scale to quantitative data) also are needed. The Australian Plant Phenomics Facility (www.plantphenomics.org.au/) is developing high-throughput phenotyping platforms for reproducibly capturing quantitative phenotypic data in parallel with environmental conditions.

Organizational Systems for Phenotypic Data. One organizational system pioneered through GRAMENE and other "-omics"-related projects (www.plantontology.org) uses hierarchical ontology for trait data, with vocabularies derived from published sources and terms appropriately defined. Input from user communities outside the basic researcher vary within trait ontology efforts. Within the international Solanaceae sequencing effort and the Solanaceae Genome Network, interaction with applied research communities is growing, and the system is proving flexible enough to account for diverse traits. Efforts are under way to ensure that ontologies are consistent with existing descriptors and have quantitative definitions. However, use of these ontologies by the community is lagging, an issue that needs to be addressed. Other initiatives include the development of Phenom-Networks (phnserver.phenome-networks.com/icis/), a web-based system to import raw data and facilitate analysis across experiments. Phenom-Networks draws its standards from the International Crop Information System (ICIS), a framework for integrated management of crop-improvement data for both individual crops and farming systems. The ICIS framework is being developed by the Consultative Group on International Agriculture (CGIAR) and has established guidelines for germplasm and data management (www.icis.cgiar.org/icis/index.php/ICIS_Concepts).

Support for Germplasm Stock Centers. The availability of germplasm (plant genetic material) linked to genetic information presented in bioinformatic resources will strongly influence both the value and audience of these resources. Germplasm housed within the National Plant Germplasm System (NPGS) is of historical interest but often does not meet the needs of breeding programs today. In contrast, immortal mapping populations (e.g., recombinant inbred lines and segmental substitution populations) may be too limited for broad inferences or may be based on accessions that are more interesting to basic scientists than those actively engaged in crop improvement. Databases designed to foster crop improvement will need to accommodate mapping populations, breeding populations and pedigrees, and germplasm accessions as defined by the user community. Permanent, long-term support to maintain a germplasm stock center for bioenergy-related species is critical.

Cross-Species Analyses

An important goal is developing databases that permit comparative analyses of gene sequences, transcript and protein abundance, phenotypes, and the relationship among these components across multiple species so that the value of genomic information can be expanded. However, the challenge of developing databases designed to be useful across species begins prior to data collection or formatting. It is critical that gene orthologs, experimental conditions, and genotypes be considered before any meaningful comparison can be achieved between any two genomic datasets. Also highly valuable would be resources for connecting gene or protein expression data and other information available in the database for one species to the most likely ortholog in other species.

A few critical data integration requirements must be considered when developing the standards and tools needed to connect data across species. These requirements include defining common terms for gene function among different species, having high-quality genome annotations with accurate depictions of gene structures, and obtaining standardized ortholog sets for navigating between genomes. In addition to comparing across species, analytical tools are needed that permit meta-analysis across experimental studies.

Defining "Data Integration"

Workshop speaker Eva Huala noted in her presentation that although "data integration" is a widely used expression, it can have different meanings depending on audience and context. For example, "data integration" can mean:

- Integration of data from many experiments of a similar type in a single species (e.g., many different microarray experiments on *Arabidopsis*).
- Integration of data from experiments of different types in a single species (e.g., gene expression, protein expression, metabolic pathways to generate a network diagram or create a summary of all data for one gene).
- Integration of data from two or more species.
- Use of an integrated dataset to extract new knowledge.

Each type of "data integration" involves different sets of problems and bottlenecks. Determining whether or not data have been integrated appropriately entails much more than simply combining data; it also involves determining whether or not useful information can be extracted from the combined data.

Standards for Interoperability

There is a perception that funding practices and cultural pressures for attaining professional recognition within research communities often encourage the development of more new tools and bioinformatic resources rather than support maintenance and improvement of existing resources. With this push to build isolated, project-specific bioinformatic resources, there is little incentive to set the standards needed to promote interoperability among these resources. User metrics, such as web statistics and literature citations, are useful for evaluating the impacts and quality of tools, databases, or datasets.

When summarizing recommendations from the Workshop on Plant Bioinformatics and Databases sponsored by the European Commission-United States (EC-US) Task Force on Plant Biotechnology Research, Doreen Ware noted several efforts for which standards development could help create a unified platform for plant genome biology:

- **Assessments of genomic tools and datasets.** Establishing periodic assessments of important genomic tools and datasets, similar to CASP (Critical Assessment of Techniques for Protein Structure Prediction), will be important for monitoring the quality of datasets and selecting the best tools for data analysis and integration. This effort will be essential for ensuring best practice and quality for reference data.

- **Genome sequence assemblies.** There is a life cycle associated with sequence datasets whereby additional improvement in the annotation for a reference genome sequence is needed even after the reference genome has been completed. With recent developments in next-generation sequencing, a standardized system should be established for evaluating the quality of sequence assemblies. Mechanisms are needed to describe the range of genome sequence models and assemblies that can now be produced, and researchers need to understand the status and quality of the genome annotations with which they are working.
- **Plant-specific ontologies.** Multiple plant-specific databases have ontologies, but there are no consistent standards among them. Coordinated efforts are needed with respect to controlled vocabularies for data collection and submission across databases, such as those used by the Plant Ontology Consortium (PO; www.plantontology.org), as well as leadership-driven efforts to generate phenotype ontologies. For phenotypes relevant to plant breeding activities, a system is needed for linking the terms used in genomic functional annotations to the phenotype terms used by breeders.
- **Curation.** For community-based curation and the curation of legacy data, there currently are no agreed-upon standards.

Knowledgebase Usability and Data Availability Issues

Long-term Sustainability of Data and Databases

Workshop participants were concerned that expiration of funding for existing databases could be problematic for sustaining the availability of important data types. This issue applies to both small boutique databases and larger community databases. Transfer of data from small, project-based databases into larger, more permanent data repositories can be difficult because of differences in schema design and scope. An additional challenge for small project-scale databases is frequent periods of unavailability due to server or network problems. Although a standard database schema (Chado) exists, it is not ideal for all purposes and has performance issues for high data volumes and usage levels. Participants also noted that getting funding for new databases currently is easier than securing continued funding for existing databases, compounding the problem of data longevity. Some participants believe the creation of new resources should continue to be the funding priority, since development of something new ensures that it will be tailored to the needs of the project. Others think that funding support needs to shift toward promoting reuse of existing resources and tools to encourage emergence of standards. Promoting reuse would require that money be made available for adapting existing tools to fit new projects, as there is always some work to be done before an existing tool or standard can be used.

Cultural Differences within the Potential User Community

The diversity of the potential Knowledgebase user community suggests that a one-size-fits-all solution may be difficult to achieve. A user's scientific culture influences how he or she views data and asks questions about the data, so different users within the plant-science community have different needs and expectations from a knowledgebase. A systems biologist, for example, needs tools to discover how a plant works. A plant breeder, however, is simply interested in predicting the phenotype that results when a particular genotype is grown in a certain environment—without really needing to know how and why the observed phenotype is produced. In this case, black box methods for predicting phenotypes may suffice. A knowledgebase therefore needs to be adaptable to the different needs of diverse users. Although many existing bioinformatics resources have focused on engaging molecular biologists, genome scientists, computational biologists, and bioinformatics specialists, more effort is needed to bridge the gap and explore the information needs of users in more-applied fields such as plant breeders and crop scientists.

User-Dependent Data Formats

Users whose daily work is focused on plant breeding or laboratory experiments want to access bulk data in relatively simple formats (e.g., CSV flat files or GFF) for further manipulation on their own computers. Some interest was expressed in portability of data or databases so that work could be performed offline (e.g., while traveling or in remote areas where internet access is slow or unavailable). Other users with a more computational focus preferred more complex data formats such as XML. Nexus format for phylogenetic data also was suggested as a good standard. Participants pointed out that certain data types (e.g., sequence and microarray data) already have well-defined standards. In general, many scientists do not want to spend time addressing format issues; they want data presented to them in an intuitive way that does not require them to become programming experts.

Education, Training, and Communication

In the life sciences, adopting informatic resources requires a user community that is educated in bioinformatics concepts, methods, and tools and is equipped with skills in computational and quantitative analytical approaches from the fields of computer sciences, statistics, and mathematics. A key problem is a lack of people with sufficient training to fully exploit the genomic information and resources available. Training the current and next generations of biologists in computational and statistical methods is a major challenge.

In the physical sciences, the computational skills required to manipulate large datasets are considered indispensable and are taught to every undergraduate and graduate student in these disciplines. Similar training in computational approaches to biology is needed at all levels, especially the undergraduate. Workshop participants specifically proposed pre- and postdoctoral cross-training fellowships in quantitative genetics, bioinformatics, and computational biology of biofuel species. For maximum impact, these fellowships should not be tied to standard research grants, where typical 3-year cycles would impede recruitment of fellows, as the hire needs to be coordinated with the duration of the grant.

Existing databases can play an important role through tools that assist self-learning (e.g., online tutorials). Although there is a need to provide tools simple and intuitive enough for those without computational training, these resources should be designed to gradually enhance understanding of underlying concepts and progressively lead the user to use the tools in more sophisticated ways. An example is a query tool that provides canned statements (in Structured Query Language or other appropriate formats) that can be altered easily by users to fit their particular needs. Eventually a user should be able to write new queries based on the knowledge gained from using and modifying the examples.

Plant Bioinformatic Efforts Relevant to Knowledgebase Development

Two ongoing bioinformatics efforts for plant biology were featured in the presentations at this workshop: the iPlant collaborative funded by the National Science Foundation (NSF) and presented by Steve Rounsley (University of Arizona) and the DOE Joint Genome Institute's Phytozome, presented by Dan Rokhsar (JGI).

The iPlant Collaborative: Cyberinfrastructure for the Plant Sciences

The NSF-supported iPlant Collaborative is an effort to develop a cyberinfrastructure that is nimble enough to address an evolving array of plant science grand challenges. According to NSF, the cyberinfrastructure is a combination of High Performance Computing (HPC), data, data analysis capabilities, and virtual organizations that also can serve as a resource for training and workforce development. The collaborative establishing iPlant includes more than 25 institutions and 45 additional researchers and continues to grow. Once the research community identifies the major problems in plant sciences, iPlant's mission is to provide the cyberinfrastructure that brings together the information needed for researchers to address these grand challenges.

iPlant's community-driven process identified two grand challenge projects that will be the focus over the next 2 years:

1. **Plant Tree of Life (iPToL).** The iPToL goal is to "build the cyberinfrastructure needed to scale up phylogenetic methods by 100-fold or more, to enable the dissemination of data associated with such large trees, and to implement scalable 'post-tree' analysis tools to foster integration of the plant tree of life with the rest of the botanical science." The largest phylogenetic tree that currently can be built is about 100-fold smaller than the number of green plants that exist. For this grand challenge, iPlant aims to design the computational approach that can be used to build a tree with 500,000 taxa in it. Using algorithms available today, the largest trees that can be built contain about 55 taxa and take about 3,000 CPU hours to construct. Some of the significant computational bottlenecks that iPToL will address will require redesigning algorithms. In addition to providing needed cyberinfrastructure, this project involves building, visualizing, and extracting data from the trees.

2. **Genotype to Phenotype (iPG2P).** The goal of the Genotype to Phenotype grand challenge is to elucidate "the relationship between plant genotypes and the resultant phenotypes in complex (e.g., non-constant) environments, one of the foremost challenges in plant biology." Although solving this grand challenge is not possible in a 2-year time frame, the project aims to help overcome the current computational and data management bottlenecks preventing researchers from even attempting to address this challenge today. Much of this effort concerns handling the different data generated

from genomics experiments (e.g., sequence, expression, metabolic, whole-plant, environmental), integrating these data, bringing in the modeling and statistical inference tools to analyze the data, visualizing the results, and providing the interfaces that researchers can use to work with their own results.

Phytozome: A DOE JGI Resource for Green Plant Comparative Genomics

The DOE Joint Genome Institute (JGI) has developed a central hub (www.phytozome.net) to provide all researchers with an interface to interact with plant genomic data in a unified way. About 20 plant genomes are included in version 5 of Phytozome, and a year from now JGI is expected to have 50 genomes of similar quality. All the genomes in Phytozome are reasonably high quality drafts, with enough data available to provide an approximation of the gene set. The genomes at Phytozome range from *Arabidopsis*, which is a highly developed, well-annotated genome, to cassava, which is a 454 draft genome that has just recently become available.

Genomes can serve as an organizing principle for much of the information emanating from modern biological studies. Looking across a phylogenetic tree of angiosperms, the timescale for their radiation is comparable to diversification of mammals (~150 million years), so the extent of diversity seen in angiosperms parallels what is seen among mammals (ranging from bats to elephants to humans). Thus, the work that has been done to compare mammal and other animal genomes indicates where comparisons of plant genomes could be in a few years. Genomes are a central axis for moving from organism to organism and seeing how different species have evolved, and certain comparisons between two different species can be useful in identifying particular kinds of candidate functional elements.

Principles Guiding Future Development of Phytozome

- Adopt open-source, community standards where possible, pulling from advanced comparative genomics already under way in vertebrates.
- Provide standardized datasets to the community. Although several versions of annotation for a genome may exist, the research community needs to agree that one version serves as the reference set at any given time.
- Take advantage of the handful of reference genomes (e.g., *Arabidopsis* and maize) that have benefited from a richer history of past research to help develop resources for the numerous new genomes that will be generated from Illumina and 454 sequencing.
- Continue to develop genome annotation assistance and browsers [e.g., JGI plant pipeline and GMOD (Generic Model Organism Database project)] using open-source community standards so that any researcher can locally set up a customized GBrowse for a particular species.
- Improve an array of features by building on existing resources:
    - "Phylogenomic" gene families (calibrated molecular divergence, synteny, molecular phylogenetic methods).
    - Comparative genomics taking advantage of VISTA and comparative tools for animal genomes.

  - Genomic diversity that builds on resources developed for human HapMap.
  - Complex queries. A guiding principle is to be able to download data in a standardized format that researchers can use in a customized way.
  - Customized analysis. GALAXY and other tool kits are built to hold data in a standardized format. Once a tool is brought into GALAXY, anyone can use it on any genome.
  - Links to TAIR, DOE Bioenergy Research Center knowledgebase efforts, iPlant, and other resources.
- Support workshops to systematically annotate the gene complement across plants.

Appendix 1: Agenda

USDA National Institute of Food and Agriculture Plant Genome, Genetics, and Breeding Project Directors' Meeting

and

Joint USDA-DOE Plant Knowledgebase Workshop

Town and Country Resort and Convention Center
San Diego, California
Friday, January 8, 2010

7:30 a.m.	Light refreshments available
7:45 – 10:00 a.m.	**Morning Session I: Plant Genome, Genetics, and Breeding** *(Pacific Salon 3)*
7:45 a.m.	Ed Kaleikau, USDA NIFA "AFRI Plant Genome, Genetics and Breeding Program"
8:00 a.m.	Phil McClean, North Dakota State University "BeanCAP – A NIFA Coordinated Agricultural Project"
8:20 a.m.	Scott Jackson, Purdue University "Genome Sequence for Common Bean"
8:30 a.m.	Gary Muehlbauer, University of Minnesota "Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement"
8:50 a.m.	Tim Close, University of California, Riverside "Advancing the Barley Genome"
9:00 a.m.	Jeff Bennetzen, University of Georgia "Development of Genomic and Genetic Tools for Foxtail Millet: Use of these Tools in the Improvement of Biomass Production for Bioenergy Crops"
9:20 a.m.	John Vogel, USDA ARS *"Brachypodium distachyon*: A New Model for the Grasses"
9:40 a.m.	Peter Bretting, USDA ARS "GRIN-Global: An International Project to Develop a Global Plant Genebank Information Management System"
10:00 – 10:30 a.m.	Break
10:30 – 12:30 p.m.	**Morning Session II: Setting the Stage for the Plant Knowledgebase Workshop: Bioinformatics Use in Advancing Plant Genomics, Genetics, and Breeding** *(Pacific Salon 3)*

10:30 a.m.	Cathy Ronning, DOE BER "Introduction to the Workshop"
10:35 a.m.	Bob Cottingham, Oak Ridge National Laboratory "DOE Systems Biology Knowledgebase"
10:50 a.m.	David Francis, Ohio State University "A Plant Breeding Perspective"
11:10 a.m.	Steve Rounsley, University of Arizona "The iPlant Collaborative"
11:30 a.m.	Eva Huala, TAIR "Leveraging *Arabidopsis* Data for Research on Other Plant Species"
11:50 a.m.	Doreen Ware, Cold Spring Harbor Laboratory "US-EC Plant Bioinformatics"
12:00 p.m.	Dan Rokhsar, Joint Genome Institute "Genomes as an 'Organizing Principle' for the Knowledgebase"
12:20 p.m.	Instructions and Move to Breakout Rooms
12:30 – 2:30 p.m.	**Working Lunch: Plant Knowledgebase Breakout Sessions and Discussion** 5 Groups; Facilitators: Rex Bernardo, Steve Knapp, Robin Buell, Lukas Mueller, and Todd Mockler *(Pacific Salons 2, 4, 5, 6, 7)*
2:30 – 2:45 p.m.	Coffee Break
2:45 – 4:00 p.m.	Report out (15 minutes for each group) *(Pacific Salon 3)*
4:00 – 4:30 p.m.	Facilitators gather to summarize and wrap up *(Pacific Salon 3)*
4:00 – 6:00 p.m.	**Poster Session** *(Golden Ballroom)*

Appendix 2: Participants and Observers

Participants

Eduard Akhunov (Kansas State U.)
Steve Baenziger (U. Nebraska)
Ali Barakat (Pennsylvania State U.)
William Barbazuk (U. Florida)
Eric Beers (Virginia Tech U.)
Jeffrey Bennetzen (U. Georgia)
Rex Bernardo (U. Minnesota)
William Berzonsky (S. Dakota State U.)
Jim Bradeen (U. Minnesota)
Charles Brummer (U. Georgia)
Marcia Buanafina (Pennsylvania State U.)
Robin Buell (Michigan State U.)
John Burke (U. Georgia)
Victor Busov (Michigan Technological U.)
Patrick Byrne (Colorado State U.)
John Carlson (Pennsylvania State U.)
Tim Close (U. California - Riverside)
Luca Comai (U. California - Davis)
Carlos Crisosto (U. California - Kearney)
Richard Cronn (USDA FS)
Thomas Davis (U. New Hampshire)
Katrien Devos (U. Georgia)
Amit Dhingra (Washington State U.)
David Douches (Michigan State U.)
Andrew Doust (Oklahoma State U.)
Jorge Dubcovsky (U. California - Davis)
Ismail Dweikat (U. Nebraska)
David Francis (Ohio State U.)
Bikram Gill (Kansas State U.)
Jim Giovannoni (Cornell U.)
Jose Gonzalez (S. Dakota State U.)
Pam Green (U. Delaware)
Maria Harrison (Cornell U.)
Patrick Hayes (Oregon State U.)
Sam Hazen (U. Massachusetts)
Eva Huala (TAIR)
Amy Iezzoni (Michigan State U.)
Eric Jackson (USDA ARS)
Scott Jackson (Purdue U.)
James Kelly (Michigan State U.)
Matias Kirst (U. Florida)
Steve Knapp (U. Georgia)
Jan Leach (Colorado State U.)
Thomas Lubberstedt (Iowa State U.)
Laura Marek (Iowa State U.)
Michael Mazourek (Cornell U.)
Phil McClean (North Dakota State U.)
Susan McCouch (Cornell U.)
Richard Michelmore (U. California - Davis)
Amit Mitra (U. Nebraska)
Todd Mockler (Oregon State U.)
Gary Muehlbauer (U. Minnesota)
Lukas Mueller (Cornell U.)
Seth Murray (Texas A & M U.)
David Neale (U. California - Davis)
Joseph Onyilagha (U. Arkansas - Pine Bluff)
Jiwan Palta (U. Wisconsin)
Cameron Peace (Washington State U.)
Zhaohua Peng (Mississippi State U.)
Andy Pereira (Virginia Tech U.)
Dan Rokshar (JGI)
Pam Ronald (U. California - Davis)
Jeffrey Ross-Ibarra (U. California - Davis)
Steve Rounsley (U. Arizona)
John Warner Scott (U. Florida)
Kevin Smith (U. Minnesota)
Carol Soderlund (U. Arizona)
David Spooner (U. Wisconsin)
Dina St. Clair (U. California - Davis)
Steve Strauss (Oregon State U.)
Christian Tobias (USDA-ARS)
Jerry Tuskan (ORNL)
Allen Van Deynze (U. California - Davis)
Richard Veilleux (Virginia Tech U.)
Wilfred Vermerris (U. Florida)
John Vogel (USDA-ARS, Albany CA)
Dong Wang (U. Nebraska)
Shizhong Xu (U. California - Riverside)
Janice Zale (U. Tennessee)
Hongyan Zhu (U. Kentucky)

Observers

Peter Bretting (USDA)
Randy Johnson (USFS)
Ed Kaleikau (USDA)
Shing Kwok (USDA)
Liang-Shiou Lin (USDA)
Gail McLean (DOE)
Jack Okamura (USDA)
Jane Silverthorne (NSF)
Sharlene Weatherwax (DOE)